ANÁLISIS DE REGRESIÓN

Edgar Acuña Fernandez
(academic.uprm.edu/eacuna)

Departamento de Matemáticas
Universidad de Puerto Rico
Recinto Universitario de Mayagüez

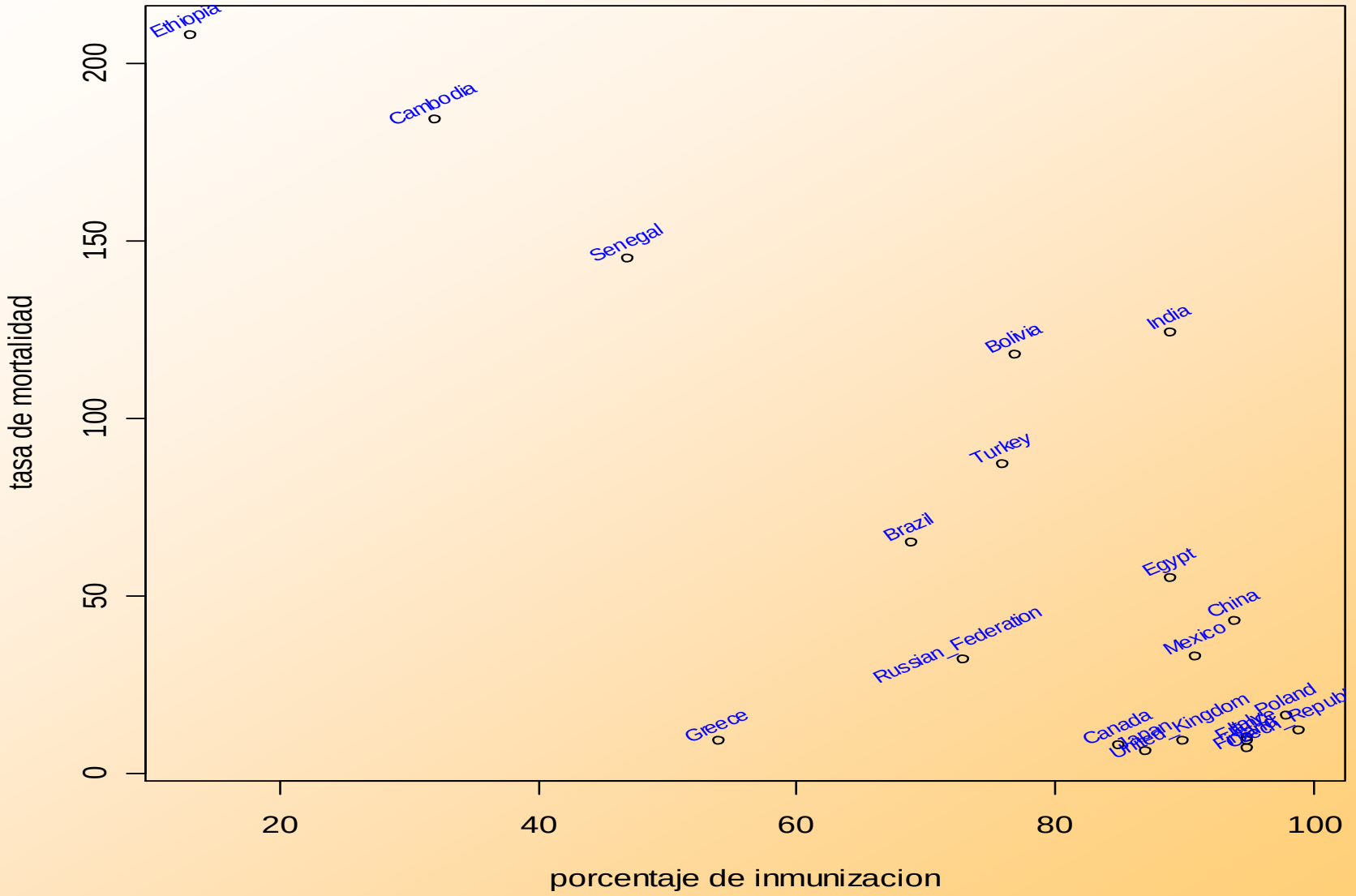
REGRESIÓN LINEAL SIMPLE

- **Regresión:** conjunto de técnicas que son usadas para establecer una relación entre una variable cuantitativa llamada *variable dependiente* y una o más variables independientes, llamadas predictoras. Estas deben ser por lo general cuantitativas, sin embargo usar predictoras que son cualitativas es permisible. Cuando hay solo una predictora se llama regresion simple.
- **Modelo de regresión.** Ecuación que representa la relación entre las variables. Cuando el modelo es lineal se llama regresion lineal.
- Para estimar la ecuación del modelo se debe tener una muestra de entrenamiento.

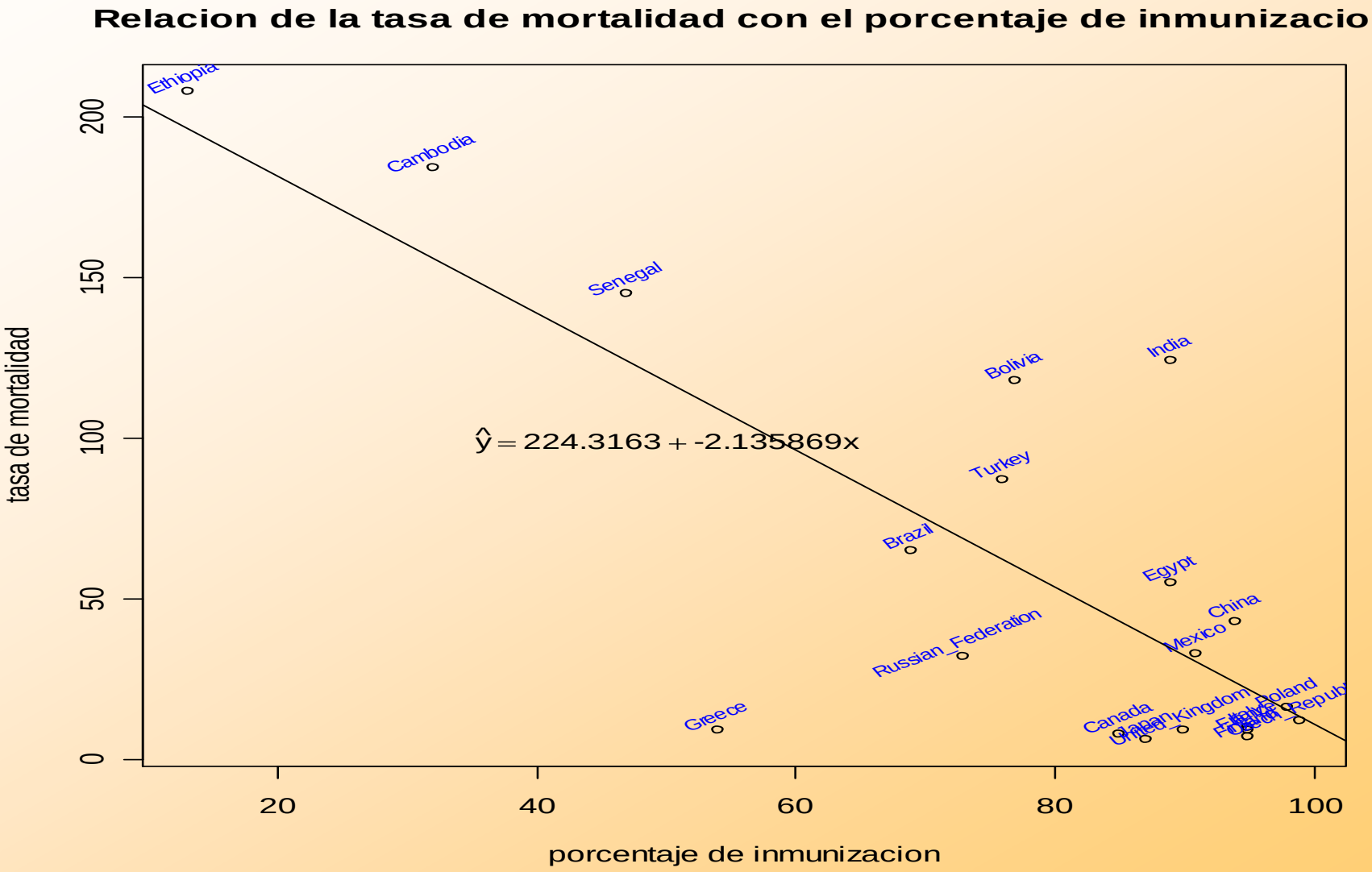
Ejemplo 1

	NACION	%INMUNIZACION	TASA_mor
1	"Bolivia"	77	118
2	"Brazil"	69	65
3	"Cambodia"	32	184
4	"Canada"	85	8
5	"China"	94	43
6	"Czech_Republic"	99	12
7	"Egypt"	89	55
8	"Ethiopia"	13	208
9	"Finland"	95	7
10	"France"	95	9
11	"Greece"	54	9
12	"India"	89	124
13	"Italy"	95	10
14	"Japan"	87	6
15	"Mexico"	91	33
16	"Poland"	98	16
17	"Russian_Federation"	73	32
18	"Senegal"	47	145
19	"Turkey"	76	87
20	"United_Kingdom"	90	9

Relacion de la tasa de mortalidad con el porcentaje de inmunizacion



Ejemplo de una linea de Regresion



1.1.1Usos del análisis de regresión

a) **Predicción**

b) **Descripción**

c) **Control**

d) **Selección de variables**

1.2 El modelo de Regresión Lineal simple

$$Y = \alpha + \beta X + \varepsilon$$

Considerando la muestra (X_i, Y_i) para i=1,…n

$$Y_i = \alpha + \beta X_i + e_i$$

- **Suposiciones del modelo:**

  La variable predictora X es no aleatoria

  Los errores e_i son variables aleatorias con media 0 y varianza constante σ^2.

  Los errores e_i y e_j (i≠j=1…,n) son independientes entre si

1.2.1Estimación de la línea de regresión usando Mínimos Cuadrados

Se debe Minimizar

$$Q(\alpha,\beta) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

Derivando parcialmente con respecto a α y β se obtiene un par de ecuaciones normales para el modelo, cuya solucion produce

$$\hat{\beta} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

O equivalentemente $\hat{\beta} = \frac{S_{xy}}{S_{xx}}$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

1.2.2 Interpretación de los coeficientes de regresión estimados

La pendiente $\hat{\beta}$ indica el cambio promedio en la variable de respuesta cuando la variable predictora aumenta en una unidad adicional.

El intercepto $\hat{\alpha}$ indica el valor promedio de la variable de respuesta cuando la variable predictora vale 0. Sin embargo carece de interpretación práctica si es irrazonable considerar que el rango de valores de x incluye a cero.

1.2.3 Propiedades de los estimadores mínimos cuadráticos de regresión

a) $\hat{\beta}$ es un estimador insegado de β. Es decir, E($\hat{\beta}$)=β

b) $\hat{\alpha}$ es un estimador insegado de α. Es decir, E($\hat{\alpha}$)=α

c) La varianza de $\hat{\beta}$ es $\frac{\sigma^2}{Sxx}$ y la de $\hat{\alpha}$ es

$$\sigma^2\left(\frac{1}{n}+\frac{\bar{x}^2}{Sxx}\right)$$

1.2.4 Distribución de los estimadores mínimos cuadráticos

Para efecto de hacer inferencia en regresión, se requiere asumir que los errores e_i , se distribuyen en forma normal e independientemente con media 0 y varianza constante σ^2. En consecuencia, también las y_i's se distribuyen normalmente con media $\alpha + \beta x_i$ y varianza σ^2.

Se puede establecer que:

$$\hat{\beta} \sim N(\beta, \frac{\sigma^2}{S_{xx}}) \qquad \hat{\alpha} \sim N(\alpha, (\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}})\sigma^2)$$

1.2.5 Propiedades de los residuales

Los residuales son las desviaciones de los valores observados de la variables de respuesta con respecto a la línea de regresión.

a) La suma de los residuales es 0. Es decir, $\sum_{i=1}^{n} r_i = 0$

b) $\sum_{i=1}^{n} r_i x_i = 0$

c) $\sum_{i=1}^{n} r_i \hat{y}_i = 0$

1.2.7 Descomposición de la suma de cuadrados total

La desviacion de un valor observado de la variable de respuesta con respecto a su media se puede escribir como:

$$(y_i - \bar{y}) = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$$

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

SST = SSE + SSR

Se puede deducir que

$$SSR = \hat{\beta}^2 \sum_{i=1}^{n}(x_i - \bar{x})^2$$

Se puede demostrar que:

$$E(SSR) = E(\hat{\beta}^2 S_{xx}) = \sigma^2 + \beta^2 S_{xx}$$

Las sumas de cuadrados son formas cuadráticas del vector aleatorio Y y por lo tanto se distribuyen como una Ji-cuadrado.

Se pueden establecer los siguientes resultados:

i) $\frac{SST}{\sigma^2} \sim \chi'^2_{(n-1)}$ (Ji-Cuadrado no central con n-1 g.l)

ii) $\frac{SSE}{\sigma^2} \sim \chi^2_{(n-2)}$ Equivalentemente $\frac{(n-2)s^2}{\sigma^2} \sim \chi^2_{(n-2)}$

iii) $\frac{SSR}{\sigma^2} \sim \chi'^2_{(1)}$ (Ji-Cuadrado no central con 1 g.l)

1.2.6 Estimación de la varianza del error

- Un estimador insesgado de σ^2 es:

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2} = \frac{\sum_{i=1}^{n} r_i^2}{n-2}$$

s^2 es tambien llamado el cuadrado medio del error (MSE)

1.2.8 El Coeficiente de Determinación R^2

Es una medida de la bondad de ajuste del modelo

$$R^2 = \frac{SSR}{SST} * 100\%$$

Un modelo de regresion con R^2 mayor o igual a 75% se puede considerar bastante aceptable.

Nota: El valor de R^2 es afectado por la presencia de valores anormales.

1.3 Inferencia en Regresion Lineal Simple

- Pruebas de hipótesis e intervalos de confianza acerca de los **coeficientes de regresión** del modelo de regresión poblacional.

- Intervalos de confianza para un valor **predicho** y para el **valor medio** de la variable de respuesta

1.3.1 Inferencia acerca de la pendiente y el intercepto usando la prueba t.

La pendiente de regresión se distribuye como una normal con media β y varianza $\frac{\sigma^2}{Sxx}$

Un intervalo de confianza del 100(1-α)% para la pendiente poblacional β es de la forma:

$$(\hat{\beta} - t_{(n-2,\alpha/2)} \frac{s}{\sqrt{Sxx}}, \hat{\beta} + t_{(n-2,\alpha/2)} \frac{s}{\sqrt{Sxx}})$$

Donde α representa el nivel de significación.

Intervalo de confianza para el intercepto α

Un intervalo de confianza del 100(1-α)% para el intercepto α de la linea de regresión poblacional es de la forma:

$$(\hat{\alpha} - t_{(n-2,\alpha/2)} s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{Sxx}}, \hat{\alpha} + t_{(n-2,\alpha/2)} s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{Sxx}})$$

Pruebas de hipótesis para la pendiente β (asumiendo que su valor es β*)

Caso I	Caso II	Caso III
Ho: $\beta=\beta^*$	Ho: $\beta=\beta^*$	Ho: $\beta=\beta^*$
Ha: $\beta<\beta^*$	Ha: $\beta\neq\beta^*$	Ha: $\beta>\beta^*$

Prueba Estadística

$$t=\frac{\hat{\beta}-\beta^*}{\frac{s}{\sqrt{Sxx}}}\sim t_{(n-2)}$$

Regla de Decisión

Rechazar Ho,	Rechazar Ho	Rechazar Ho
si $t_{cal}<-t(\alpha,n-2)$	si $\lvert t_{cal}\rvert>t(\alpha/2,n-2)$	si $t_{cal}>t(\alpha,n-2)$

*Un “P-value” cercano a cero sugiere rechazar la hipótesis nula.

1.3.2 El análisis de varianza para regresión lineal simple

El análisis de varianza para regresión consiste en descomponer la variación total de la variable de respuesta en varias partes llamadas **fuentes de variación**.

La división de la suma de cuadrados por sus grados de libertad es llamada **cuadrado medio**.
Así se tienen tres cuadrados medios.

Cuadrado Medio de Regresión=MSR=SSR/1
Cuadrado Medio del Error= MSE=SSE/(n-2)
Cuadrado Medio del Total=MST=SST/(n-1)

Tabla de Análisis de Varianza

Fuente de Variación	g.l.	Sumas de Cuadrados	Cuadrados Medios	F
Debido a la Regresion	1	SSR	MSR=SSR/1	$\frac{MSR}{MSE}$
Error	n-2	SSE	MSE=SSE/(n-2)	
Total	n-1	SST		

Se rechazaría la hipótesis nula Ho:$\beta=0$ si el “P-value” de la prueba de F es menor de 0.05

Intervalo de confianza para el valor medio de la variable de respuesta e Intervalo de Predicción

Queremos predecir el valor medio de las Y para un valor x_0 de la variable predictora x.

$$E(Y / x = x_0) = \alpha + \beta x_0$$

El estimador natural es $\hat{Y}_o = \hat{\alpha} + \hat{\beta} x_o$ Como las Y's se distribuyen normalmente, entonces también $\hat{Y}_o$ se distribuye normalmente con media E(Y/X=xo)y varianza igual a:

$$Var(\hat{Y}_0) = \sigma^2 (\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{Sxx})$$

Intervalo de confianza (cont)

Un intervalo de confianza del 100(1-α)% para el **valor medio de las y's** dado que x=x_0 es de la forma:

$$\hat{\alpha} + \hat{\beta}x_0 \pm t_{(\alpha/2,n-2)} s\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{Sxx}}$$

Trabajando con la diferencia $Y_0 - \hat{Y}_0$ se tiene

$$E(Y_0 - \hat{Y}_0) = 0 \qquad Var(Y_0 - \hat{Y}_0) = \sigma^2(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{Sxx})$$

Luego el **intervalo de predicción para un valor individual** de Y dado x=x_0

es de la forma $$\hat{\alpha} + \hat{\beta}x_0 \pm t_{(\alpha/2,n-2)} s\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{Sxx}}$$

1.4 El Coeficiente de Correlación

Mide el grado de asociación lineal entre las variables X y Y y se define como:

$$\rho = \frac{Cov(X,Y)}{\sigma_x \sigma_y}$$

a) $-1 \le \rho \le 1$

b) Si la regresion de Y sobre X es lineal, Esto es, $E(Y/X) = \alpha + \beta x$ entonces: $\beta = \rho \frac{\sigma_y}{\sigma_x}$ y $\alpha = \mu_y - \beta \mu_x$

c) La varianza condicional de las Y dado X, está dado por

$$\sigma^2_{y/x} = \sigma^2_y (1 - \rho^2)$$

Si $\rho = \pm 1$ entonces $\sigma^2_{y/x} = 0$ (perfecta relación lineal).

Coeficiente de correlación muestral

Considerando una muestra de *n* pares (x_i, y_i)

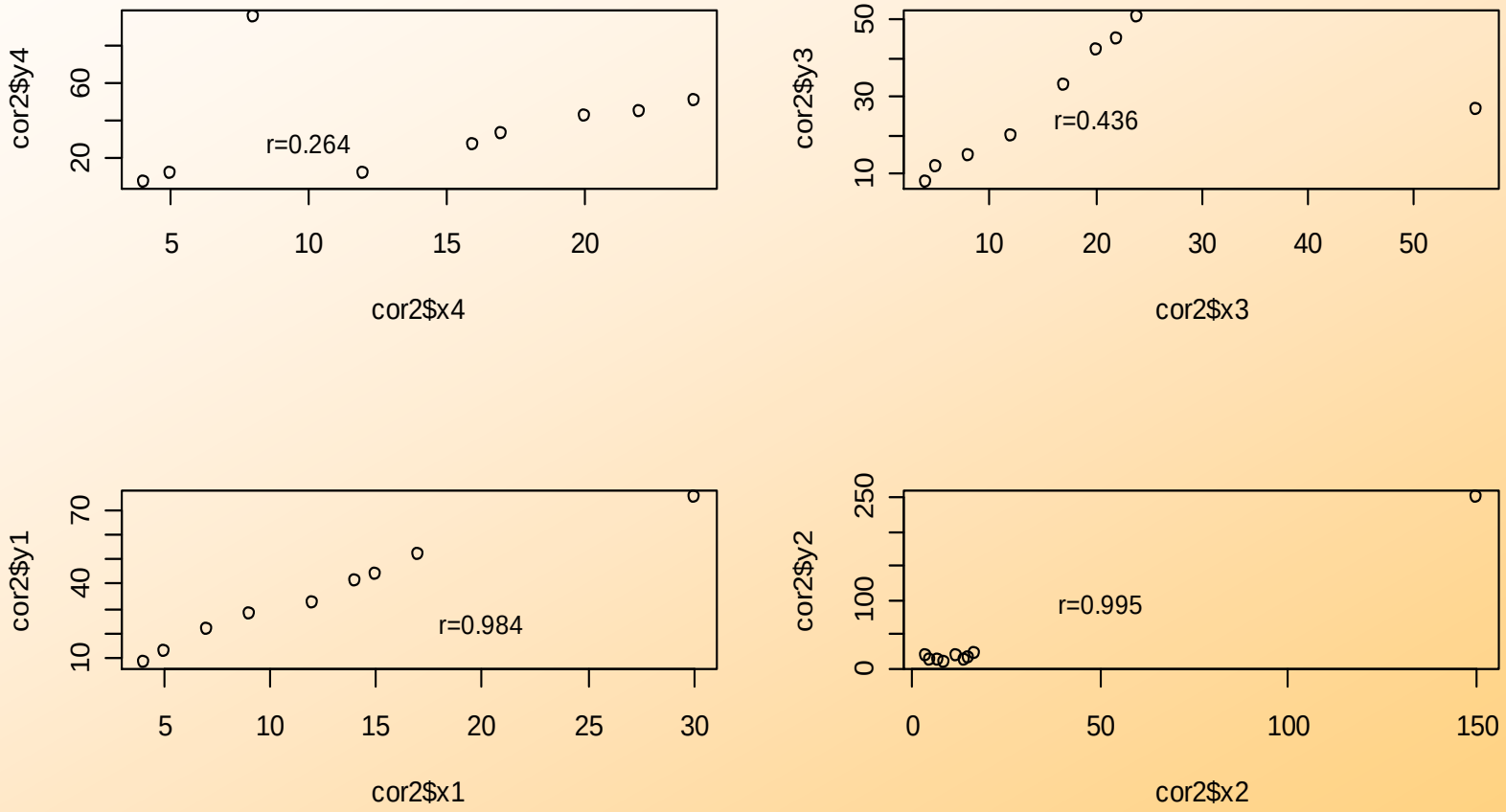
$$r = \frac{Sxy}{\sqrt{SxxSyy}}$$

Notar que:

$$r = \hat{\beta}\sqrt{\frac{Sxx}{Syy}} \qquad r^2 = \frac{\hat{\beta}^2 Sxx}{Syy} = \frac{SSR}{SST}$$

El cuadrado del coeficiente de correlación es igual al coeficiente de determinación.

Efecto de outliers en la correlacion



1.5 Análisis de residuales

Los residuales, son estimaciones de los errores del modelo y sirven para establecer si las suposiciones del modelo se cumplen y para explorar el porqué de un mal ajuste del modelo. Podemos ver:

- Si la distribución de los errores es normal y sin “outliers”.
- Si la varianza de los errores es constante y si se requieren transformaciones de las variables.
- Si la relación entre las variables es efectivamente lineal o presenta algún tipo de curvatura
- Si hay dependencia de los errores, especialmente en el caso de que la variable predictora sea tiempo.

Tipos de residuales

i) Residual Estandarizado, se divide el residual entre la desviación estándar del error. Es decir,

$$\frac{y_i - \hat{y}_i}{s}$$

ii) Residual Estudentizado, se divide el residual entre su desviación estándar estimada. Es decir,

$$\frac{y_i - \hat{y}_i}{s\sqrt{(1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{Sxx})}}$$

1.5.1 Cotejando normalidad de los errores y detectando outliers

La normalidad de los errores es un requisito indispensable para que tengan validez las pruebas estadísticas de *t* y F que se usan en regresión.

La manera más fácil es usando gráficas tales como: histogramas,

“stem-and-leaf” o “Boxplots”.

El plot de Normalidad, plotea los residuales versus los scores normales (valores que se esperarían si existiera normalidad).

La libreria car tiene una function qqPLot que hace el plot de Normalidad incluyendo bandas de confianza calculadas usando bootstrap

1.5.2 Cotejando que la varianza sea constante

Se plotea los residuales estandarizados versus los valores ajustados o versus la variable predictora X.

Si los puntos del plot caen en una franja horizontal alrededor de 0 entonces la varianza es constante.

Si los puntos siguen algún patrón entonces se dice que la varianza no es constante.

Nota: Se debe tener cuidado con la presencia de outliers.

La libreria car tiene una function ncvTest que usa la prueba de Breusch–Pagan para varianza constant que usa una prueba de Chisquare. La Hipotesis Nula es que la varianza es constante

1.5.3 Cotejando si los errores estan correlacionados.

Cuando la variable predictora es tiempo, puede ocurrir que los errores esten correlacionados secuencialmente entre si.

Prueba de Durbin-Watson, mide el grado de correlación de un error con el anterior y el posterior a él.

Estadístico

$$D = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$$

D varía entre 0 y 4.

Si D esta cerca de 0 los errores están correlacionados positivamente.

Si D está cerca de 4 entonces la correlación es negativa.

La distribución de D es simétrica con respecto a 2. Así que un valor de D cercano a 2 indica que no hay correlación de los errores.

Errores autocorrelacionados (2)

En el ejemplo resulta
el estadistico Durbin Watson de la regresion lineal es= 2.678912
Algo cerca de 2 lo cual significa que no habria correlacion de los errores

La librería car tiene una función durbinWatsonTest que prueba si hay autocorrelación o no

```
 lag Autocorrelation D-W Statistic p-value
   1      -0.4069361      2.678912   0.108
 Alternative hypothesis: rho != 0
```

Como el P-value 0.108 es mayor que .05 se concluye que no se rechaza la hipotesis nula y se concluye que no hay correlacion de los errores